Digital Archive Issues from the Perspective of an Earth Science Data Producer

Bruce R. Barkstrom

Atmospheric Sciences Division

NASA Langley Research Center

Hampton, VA 23681-0001

Table of Contents for this document

---

1 Introduction

1. Earth science data have several unique characteristics that play a role in that data's archival:

- Earth science data are always obtained with an underlying structure of spatial and temporal sampling.
- Earth science data have physical meaning, including ranges of acceptable values and physical consistency constraints.
- Earth science data appear as the results of production methods that often involve complex algorithms and tangled webs of production and validation.

2. Each of these characteristics influences the way data producers generate their data and the way users access and retrieve it. Of course, the perspective of the data producers is far from identical with the perspective of the many user communities. As we explore these perspectives and their implications in this paper, we will be reminded that digital archives have a long-term view of their data, but that they also need to be in intimate contact with the living communities of producers and users for that data.

3. We anticipate that the producer and user communities will experience significant changes in the way they produce data, store it, search for it, and use it as they gain experience in working with data that are linked in the fashion of 'hypertext'. The WWW pages and links that move us from web site to web site show how the familiar 'Table of Contents' and 'Index' expand when computers power these 'search engines' for books. Eventually the influence of these pages and links will reach into the way data producers package their 'data products'. When we arrive at that part of the next millenium, fluid forms of data packaging will fit naturally into the world of the digital archive. However, we still have the pangs of childhood and adolescence to endure before this field is mature.

4. In the body of this paper, we begin with the views of the data producers. We are particularly interested in the data structures that producers currently create. These structures often reveal a great deal about the scientific producer's 'world view'. They also reveal the constraints that current technology imposes on data production. For example, most of the data producers in the first EOS missions produce 'files'. One of the reasons they do so is that database approaches still raise serious concerns about data throughput. An additional concern arises from the complexity of data production and configuration management. We illustrate this complexity by showing how

these file structures provide a useful ‘tree’ to index the data in the archives. Data producers have to introduce special branches in these trees to account for versions of data products that arise from validating scientific data. We also recognize that this tree structure for files and file collections corresponds to a ‘Table of Contents’ for a particular data producer’s products. Metadata that aggregates data values in the data files then plays the role of an ‘Index’ that allows users to search through data in a ‘random access’ mode.

5. The second section of the body of this text considers how users view what the data producers create. Users often want to search for data in ways that are quite different from the paths that data producers use. We first consider the diversity of user communities that may access Earth science data. From this discussion, we see that users often adopt an ‘object-oriented’ frame of mind, a ‘retail’ approach to data. Data producers, on the other hand, tend to look at data from a ‘wholesaler’ point of view that emphasizes uniform blocks of data that do not have many external differences. When the archivist needs to mediate between these two points of view, he may want to use several different approaches to building searchable structures. He might include producer-provided statistical summaries of the data product files. He might use secondary indices to the files. Independent investigators could build such indices. We are probably just at the beginning of an era in which this kind of ‘data scholarship’ will markedly expand the usefulness of scientific data archives.

6. In the third section of this text, we consider the role of digital archive centers for scientific data. Before we discuss this role, we consider that the divergence between data producer views and data user views naturally leads us to consider subsetting data in files and then combining the subsets into new supersets. In print libraries, such an approach would be equivalent to allowing users to Xerox individual pages, then to excerpt the material from the pages, and finally, to create new scholarly works from interpretations of the excerpts. With subsetting and supersetting of the basic data, scientists can produce entirely new interpretations of natural phenomena. Of course, these new interpretations need peer review and extended discussion by scientists before they provide the scientific community accepts them. The archive centers that contain the data may help by serving as guides and facilitators for this discussion and consensus building activity. In order to fulfill this role, these centers need to

- Provide documentation about the data
- Provide new mechanisms for secondary indexing and other kinds of data searches
- Develop, in conjunction with scientific researchers, ways in which old data sets can be made useful to the community

7. This perspective suggests that in the long run, digital archive centers will become centers for ‘scholarly access’ to scientific data. It also suggests that the current notion of placing ‘used’ data in ‘long-term archives’ (which often appear to be viewed as ‘write-once, read-never’ data sinks in a managerial context) is incorrect. A much more useful notion is to view digital archive centers as specialized research libraries that are one component of a community of data scholars.

2 A Producer Perspective on Earth Science Data

2.1 Data Producers as Members of a Scientific Community

8. From the perspective of this paper, a data producer is a scientist (or a scientific team) who agrees to produce data for peer-reviewed scientific work. Because a critical component of scientific work is being able to reproduce

results, data producers provide their data to institutions we will call data centers. In previous times, data producers created their own data centers and distributed their data to other researchers. However, it has been helpful to have data centers that can deal with data distribution, documentation, and preservation. In some cases, data centers also partner with data producers to generate scientific data products.

9. Clearly, the scientists who generate data products are familiar with the instrumentation that generates the raw data. In addition, these scientists need to be familiar with the algorithms that convert raw data into useful scientific information. Given these skills and knowledge, data producers form a relatively homogeneous reference community for scientific data. In other words, scientific data are created within a community familiar with the disciplinary basis for the scientific investigations that serve as the initial justification for collecting the data.

2.2 Some Unique Characteristics of Scientific Data

10. If a data center was dealing with financial data, it would need computers to collect, manipulate, and store these data. In the present environment, such financial data are often generated in discrete transactions from sites such as Automated Teller Machines or Point of Sale terminals in stores. From these discrete sites, they are transmitted to more centralized locations for ingesting into databases. There, the computers create statistical summaries and provide the results to programs that compare the transaction summaries with results from programs that model the financial flows within the firm and within the economy as a whole.

11. In what ways do scientific data differ from financial data? We can identify several unique characteristics of scientific data, and of Earth science data in particular:

- First, scientific data come from instruments that depend on physical principles that may be better understood than consumer psychology or the forces that drive wholesale economic transactions. These physical principles constrain the possible values from instruments and create strong relationships between different kinds of measurements.
- Second, Earth science data always depend on a sampling of space and time. In other words, each measurement comes from a very particular region of three-dimensional space and a well-defined interval of time.
- Third, scientific data often come in the form of a continuous stream of numbers. Some data are discretely sampled in time and space. However, many Earth science data sources make measurements constantly. This fact is particularly important for data from satellites that currently provide the richest sources of information for the Earth sciences.

Let us discuss these characteristics in more detail. They govern much of what data producers do.

2.3 Spatial and Temporal Sampling for Earth (or Space) Science Data

12. The spatial and temporal coordinates that underlie Earth science data are critical to these scientific data values. In a few cases (mainly for in-situ measurements), instruments provide samples of the environment at particular points in space. However, in remote sensing, the data points sample volumes of the Earth's atmosphere or areas on its surface. For example, in remote sensing with a satellite imager, the data value in a given pixel, *m*,

is related to the radiance, *I*, leaving the top of the atmosphere at a latitude, λ, and longitude, ϕ, through a *Point Spread Function* (PSF). The imager has a coordinate system related to its optical axis, the position of the satellite (also in latitude and longitude), and the satellite's attitude. Without going through a detailed derivation, we can write the relationship as an integral:

$$m(\lambda, \phi, t) = \int d\lambda \int d\phi \, PSF(\lambda, \phi | \lambda, \phi) \, I(\lambda, \phi, t)$$

Even if a data producer records the latitude, longitude, and time of the data value, a data user may still need information about the PSF in order to determine how much the data taking process has smeared the underlying distribution of radiance.

13. For the same reason, data users may need very detailed information about the spatial and temporal sampling pattern of the instruments that collects the data. While data producers often supply illustrations of sampling patterns, they must supply quantitative descriptions of them as well. Even a simple push-broom sensor that samples from limb-to-limb is likely to have samples that are closer together at nadir (in distance on the Earth's surface) than they are near the limb. A microwave imager may use a conical scan, in which the data points are uniformly spaced around the scan. A user that wants to compare the push-broom imager data with the conical scanning microwave data will have to understand the geometry of both instruments. The need for this kind of detailed and precise sampling information is one of the distinguishing characteristics of scientific data.

2.4 The Influence of the Data Production System Architecture

14. Data producers tend to design production system architectures that make the data they work with as uniform as possible. Thus, they tend to work with uniform temporal intervals (hours, days, or months) or uniform spatial chunks (the entire globe, latitudinal profiles that are averages over all longitudes, equal-angle grids or equal-area grids). They often design their production processes so that data enters these processes in uniform time intervals even though the data have complex spatial sampling patterns within the time interval. Later in the processing, the same data producer may rearrange the data into a more uniform spatial structure that has a time series for each spatial region – particularly if the data come from an instrument that creates a more-or-less continuous stream of measurements.

2.5 The Spatial and Temporal Structures Underlying Earth Science Data

15. Figure 1 is a schematic representation of the underlying spatial and temporal structure for Earth science data. Some in-situ instruments take data from a fixed point on the Earth. This kind of sampling would appear in figure 1 as a straight line that goes from the front surface to the back, parallel to the time axis. Geostationary satellite instruments typically collect data from a circular cylinder in this figure. The center of the cylinder is located directly under the satellite. Its axis extends back through the volume, parallel to the time axis. Low-Earth orbiter instruments weave a sideways, 'S'-shaped swath through this sampling space.

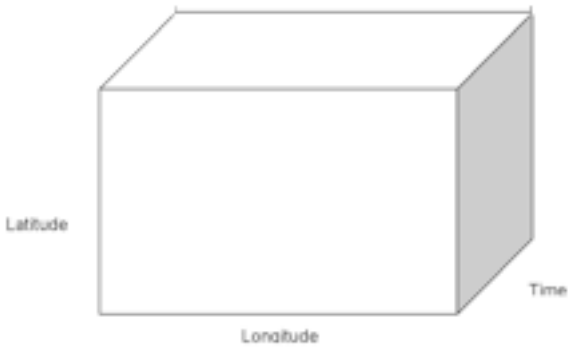


Figure 1. Projected Spatial and Temporal Coordinates for Earth Science Data. For many kinds of Earth science data, it is useful to think of the data obtained by various sensors as having attached variables that give spatial location and time for each data value. In complete generality, there are three spatial variables (latitude, longitude, and altitude) as well as time. As this schematic figure illustrates, we often project the three spatial coordinates onto two.

2.6 Earth Science Data 'File' (or Relation) Schemas

16. Data producers and archivists are not entirely consistent in recording these underlying spatial and temporal sampling structures. One factor that influences both of these groups is the expense of storing large amounts of data. In products that contain gridded data, the structure is regular enough that data producers may not include the grid boundaries in the data product. Often, they feel that they can reference the grid in documentation. They expect users to work with the data just as well as they could if they put the grid directly into the file. In other cases, the data producer will embed sample locations within the data products themselves.

17. Figure 2 illustrates a 'canonical' way of thinking about how data may be incorporated into the 'files' within a data product. We assume that in this representation, the data producer arranges the data values into records. The first four values in a record provide the centroid of the space and time sampling for the data values. The data values after these first four are then the measurement values taken within this sampling volume. Noting that data producers and archivists desire 'self-documenting' files, we also show that each file contains a second collection of records that 'annotate' the data values. As figure 2 shows, our suggestion for this annotation provides a *Name* for each field, a *Description* of it, the *Units* of that field's measurements, *Bad Value Definitions*, and similar kinds of information.

18. The form shown in figure 2 is intended to be isomorphic with the layout of data fields in a conventional database. If a data producer wanted to use this kind of software, he would design the tables similar to the rows of records and the columns of fields. A few of the data producers for the Earth Observing System have moved in this direction.

19. More commonly, data producers rearrange the sequence of fields and records. They may choose to build blocks of storage that contain only one field. The file then consists of sequences of homogeneous blocks of data values. Often, data producers do not include all of the fields identified in light gray in this figure. For example, if a data producer is creating files that contain images, he might choose to leave out the time field and the altitude field. He might embed samples of latitude and longitude every tenth line and for every tenth pixel in the image, similar to the structure NOAA uses for providing data from the Advanced Very High Resolution Radiometer (AVHRR). Alternatively, if there are many spectral bands for a very high data rate instrument, the data producer might put the latitude and longitude of each pixel in a separate file that users access when they need that

information.

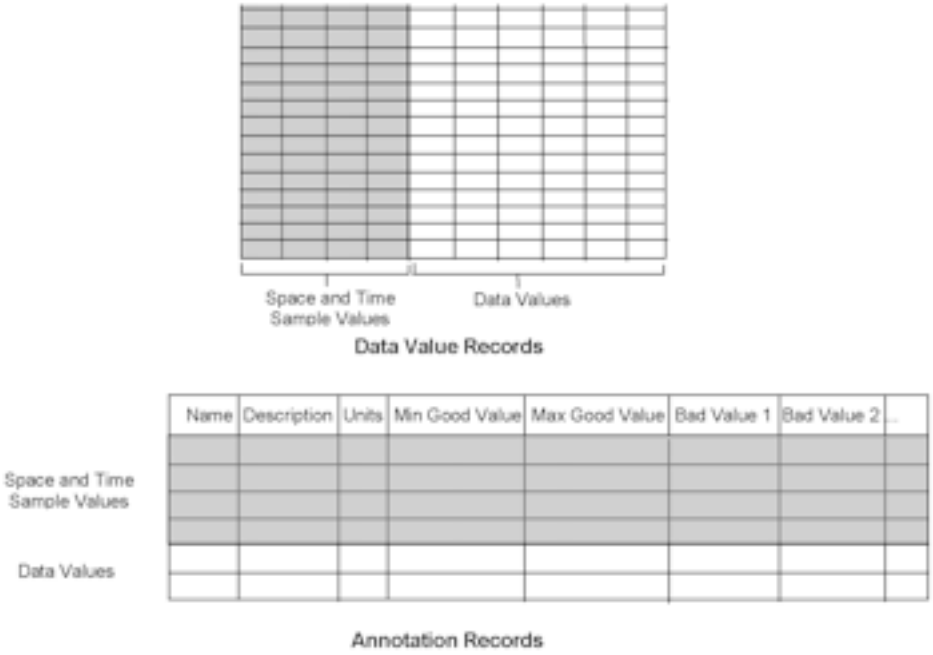


Figure 2. Canonical Treatment of Data and Annotation Structures within a 'File'. In this schematic illustration of the way data appear within a single 'file', we show records of data values and records for annotating the fields within the data value records. As we show the data structure, data values collected by a particular sampling of space and time are stored in records of successive fields. Each field has a name, a description, units, etc., as we show in the "Annotation" Records. A complete file would contain at least these two kinds of records if it were to be completely self-documenting. In practice, scientific files may treat the space and time sample values implicitly so that they are not recorded in the file at all. We display the Data Value Records and the Annotation Records in a form appropriate for database tables. In practice, data producers may group the fields and annotations in other sequences than the one we show here.

20. Scientific data from real instruments also contains 'bad values'. Sometimes the data collection system doesn't work and loses data. At other times, the instrumentation records incorrect values. In either event, data producers create quality assurance logic in their software and institute quality assurance procedures that help their teams identify bad values. The indicators of bad values are sometimes 'flag' fields and are sometimes encoded as particular data values. The canonical data structure in figure 2 supports either method of communicating predefined bad values.

2.7 Data Producer Configuration Management Complexities

21. The data producer that creates the data and the data center that stores it for distribution have the responsibility for ensuring that the data fields and the ties to the spatial and temporal sampling are adequately documented for users. In addition, the producer and the data center need to pay careful attention to configuration management of scientific data. This task can become quite complex as we show in the next few figures and paragraphs.

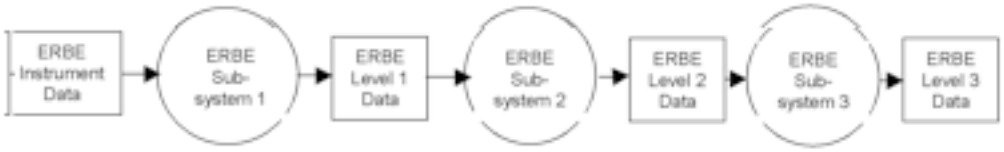
22. Data users are often fooled about the simplicity of data production because producers present simple diagrams showing how archival data products are connected with the algorithms that ingest one form of data to create another. The top diagram in figure 3a shows such a simplistic view for production of data by the Earth

Radiation Budget Experiment (ERBE). In this diagram, data from the ERBE instruments enters on the left. The first process converts the raw instrument data into instantaneous fluxes at the top of the Earth's atmosphere. The second process averages the instantaneous fluxes to produce monthly averages. However, this linear depiction is much too simple to describe the actual dependence of the archival products upon the ancillary data that the ERBE production team has to supply.

23. The lower diagram in figure 3a shows these additional kinds of files. The first process has to have calibration coefficients and satellite ephemeris data. The second process has to have Angular Distribution Models that enter directly into the conversion from radiances to fluxes. This process also needs instrument spectral characteristics and a map of the underlying surfaces that cover the Earth (oceans are different than deserts, etc.). The monthly averaging process needs to account for the systematic diurnal pattern of desert surface heating during the sunlight hours, as well as the variation of albedo over the course of the day. These technical details are not the primary concern of an archivist. However, the complexity of this topology is an important characteristic of scientific data. The lower diagram in this figure has a considerably more complex topology than the advertising diagram in the upper portion.

24. As the astute reader might expect, even the lower diagram in figure 3a does not convey the true complexity of production topology. Figure 3b shows some of that complexity for producing the monthly average, albeit assuming that a month has just two days. Of course a real month has about thirty days, which means that the daily parts of the process would be repeated thirty times. To be completely correct, we would have to include the last day of the previous month and the first day of the following month. In addition, the ERBE monthly production had to include options that process data from only one satellite or as many as three. We leave the appropriate diagram connecting files and processes to the reader's imagination.

25. The connectivity we illustrate directly concerns the data center and the serious data user. It determines the traceability or genealogy of data in a particular data product. It also affects how producers label versions of data sets. Data producers spend a great deal of their time validating data after they start producing it. A substantial part of that work involves finding unexpected discrepancies between data from one source and data from other sources. When such a discrepancy is discovered and verified, the data producer teams have to identify the source and correct the problem. Sometimes the problem lies in the algorithms that underlie the production processes. At other times, the problem lies in the input data. Once the producer identifies the cause and corrects it, he has to decide how to remove the problem from the data in the data center. Occasionally, the problem is small enough that the producer can simply note it for the users to take into account when they have a unique data use. More likely, some of the data in the data center will have to be reprocessed. After reprocessing, the data user will encounter the problem of multiple versions of the data.



Linear Simplification of Production Topology

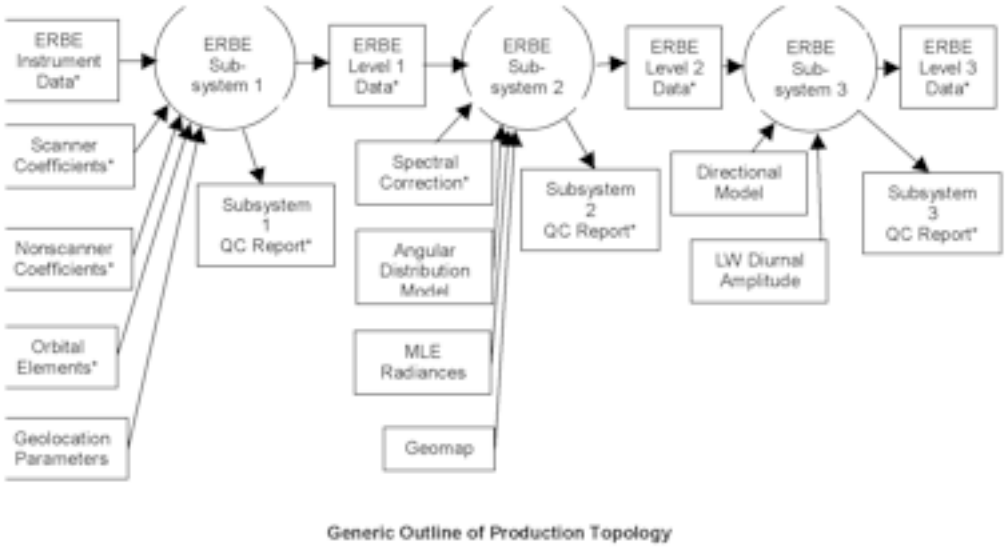


Figure 3a. Production Topology Issues – Simplistic Connectivity of 'Files' Used in Data Production. From an external view, it is often useful to present an extremely simple view of the relationship between data product files and the production processes. The upper view shows such a simplification, in which we remove all of the data 'files' except those that are the 'archival' data product files. The lower view shows the generic data products that enter the production 'jobs'. Data 'files' are shown as rectangles; processes that ingest and create data are shown as circles.

26. Figure 3c illustrates the feedback processes involved in scientific data validation. In the ERBE measurements that we illustrate here, each satellite carried both a scanning instrument and a non-scanning one. Both instrument types also carried internal and solar calibration sources. These instruments flew on several different satellites whose sampling of the top of the atmosphere overlapped at orbit intersections. Each of these calibration sources and intercomparison opportunities offered a separate validation intercomparison whose results could feed back on the coefficients that produced the data. For example, if the solar calibration targets suggested that the non-scanner instrument changed gain, then the team would change the non-scanner calibration coefficients. As we can see from the figure, multiple validation opportunities introduce complex feedback loops – at the same time that they increase the certainty of the data.

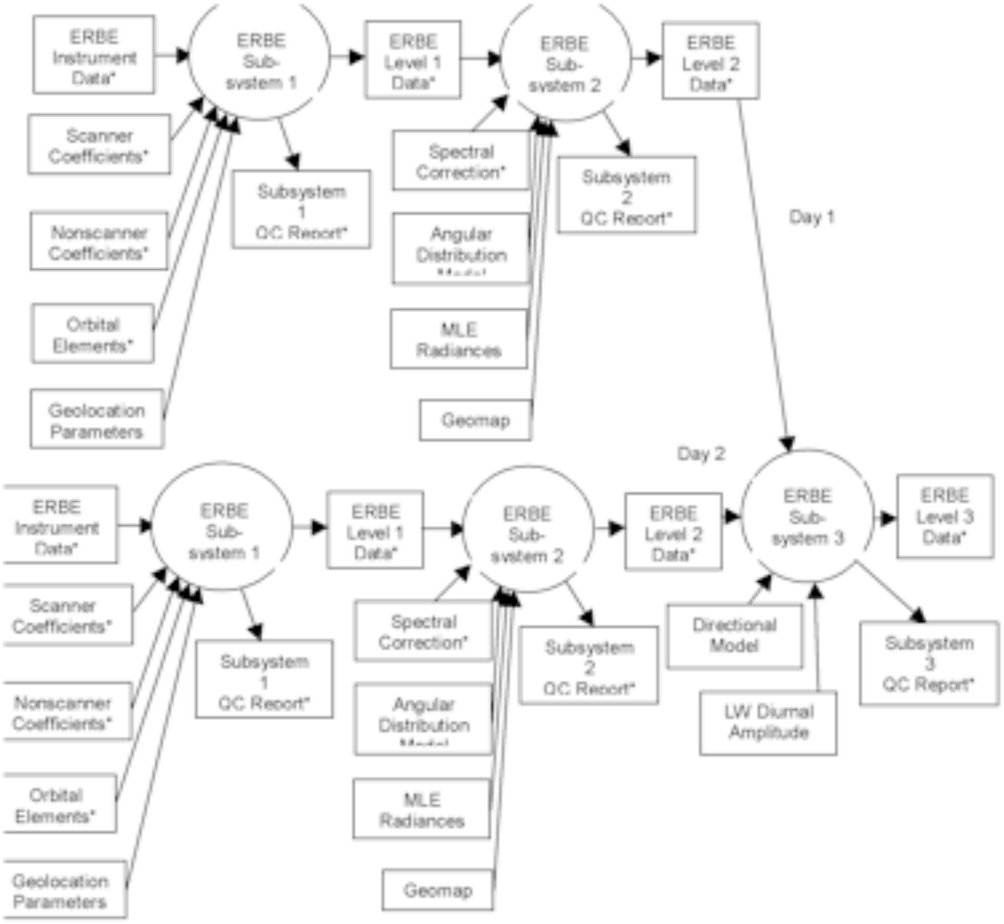


Figure 3b. Production Topology Issues – Genealogical Complexity. This figure illustrates a partial expansion of the ‘file-process’ topology that represents an actual production run which creates a monthly average data product from a month of input data. For careful archival work, the final data products need a configuration management system that will allow a researcher to track the genealogy of the antecedent data used by the production process, as well as the algorithms that created the final data products.

27. Many Earth science data producers work to produce data products that have long-term consistency. For this purpose, they may need to generate global data products for several months in a year before they can evaluate whether or not the data need further refinement. The ERBE data provide an excellent example of this requirement. Obtaining a net radiation balance that is close to zero over an annual cycle is a very stringent requirement. The ERBE science team was not satisfied with their data until they had computed the net radiation balance with a year of data to see that the observed imbalance was relatively small (which it was). Many other groups make such long-term, large area consistency checks part of their validation effort. From a user perspective, it is easy to talk of ‘eager’ and ‘lazy’ approaches to data production. From a data producer perspective, either of these approaches is overly simplistic. When they have to do serious validation work, data producers would probably prefer to describe their approach as ‘considered’.

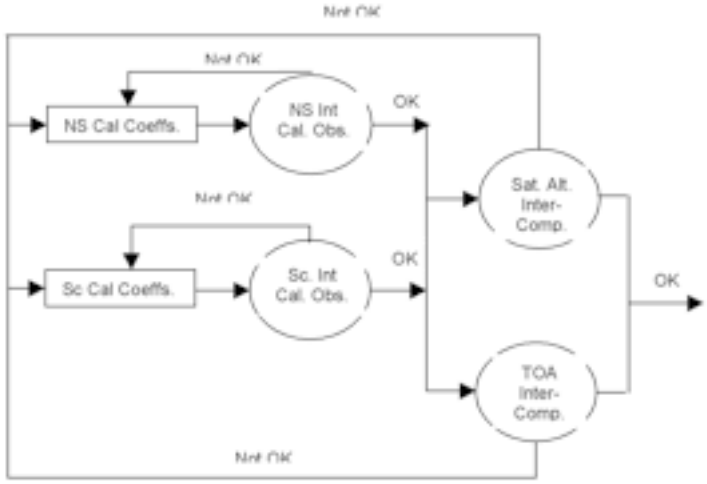


Figure 3c. Production Topology Issues – Versioning. This figure illustrates a partial expansion of the validation process that data producers use to reduce uncertainties in their final data products. The rectangles indicate the coefficient files the producers use to generate the calibrated data. Each change in these files generates a new Data Set Version in the hierarchy we describe below.

2.8 The Topology of Earth Science Data Inventories

28. Data producers generate collections of files using process-product topologies that contain complex webs of coefficient and algorithm genealogies. These webs are not random, with untraceable weavings of relationships. Rather, they are like tapestries with highly regular patterns. Data producers create groups of files that are very similar to one another, particularly for satellite data. Within one of these groups, the fields within the records have the same sequence and data types. In figure 4, we call this top level of homogeneity the *Data Product* level. A Data Product consists of *Data Sets*, in which the files contain data from a homogeneous collection of sources (a single satellite, a single in-situ data source). For ERBE we would have some Data Sets that contained only data from the Earth Radiation Budget Satellite (ERBS), others from NOAA-9 and some that contained both. As a science team proceeds through validation, they modify either the algorithms or coefficients. These variations introduce *Data Set Versions* within the Data Sets. In practice, we may need to introduce an additional configuration version to produce a unique index for each file.

29. The data, files, and file collections form a ‘natural’ hierarchy for data producers, a tree structure. Figure 4 illustrates this structure. Assuming that a data producer does not put the same data record twice into a data center, this tree provides a *unique* location for each data record. We also note that in many cases, we can place the files within a Data Set Version into a sequence ordered by the starting time of the data the file contains. In more formal terms, we might say that time and space sequences often provide a unique indexing for the possible file opportunities within a Data Set. As the files are created within a Data Set Version, the data center can relate the ‘file opportunity index’ to the file position in the version sequence.

Figure 4. Data, Files, and File Collections as a Tree. This hierarchy describes a standard organization of Earth science data that is common in the EOS Distributed Active Archive centers. Most of the data is kept in *files* made of records. The records consist of data values that are typically a 'float' data type, i.e., 4-byte floating-point numbers. *Data set versions* are intended to be "as homogeneous as possible" in their contents, so that both processing algorithms and input coefficient changes are minimized within the files in this grouping. *Data sets* are groups of data set versions, in which the dominant difference from data set to data set is the time and space sampling. *Data products* are collections of data sets. Data products are likely to be fairly uniform in content and structure. Data files may contain several record types, as well as appropriate documentation and simple metadata. These metadata components do not appear in this figure.

30. The material we have covered briefly here should convey some of the perspective that data producers bring to the problems of data production and archival. Data producers are not insensitive to users. However, the users that are most important to them are usually scientists with interests and experiences similar to the data producer. Production performance becomes paramount. These influences lead data producers to treat the data for which they are responsible in terms of relatively homogeneous 'chunks' of data that are often easiest to order in a time sequence or in gridded spatial structures. Data producers are also more likely to be sensitive to configuration issues than are data users. In the section that follows, we consider the other point of view.

3 Some Thoughts on the User Perspective

3.1 Science Data User Communities

31. In contrast with the relative homogeneity of the data producer's community, users come from diverse communities:

- Discipline-based scientific researchers similar to the data producer researchers
- Interdisciplinary researchers – who have special needs for careful documentation of spatial and temporal sampling, instrument calibration, and data production algorithms

- Commercial users – who have special needs for data produced very shortly after collection, but with less careful validation
- Educational users – who need special curricular background material and examples
- General public users – who seek information and novelty, but need good, narrative interpretations

32. As we move from the highly specialized world of the scientific researcher to the diverse inclinations and background of the general public, it becomes increasingly important to provide a support infrastructure. We expect researchers to be familiar with long words and precise understanding of the meaning of data annotation. The general public grows impatient with long words and long definitions. Regardless, none of these communities wants to wait for data.

3.2 Spatial and Temporal Structure Needs of Different Users

33. The divergence among the user communities forces data centers to take several different approaches when they present the spatial and temporal structure of data to data users. Researchers working in a particular scientific discipline receive grounding in the spatial and temporal structures of their discipline's data as part of their education. Concise tables or documentation written by other researchers probably suffice for researchers. Researchers doing interdisciplinary work are likely to have to deal with a variety of conventions. A common documentation format for describing coordinate systems and data formats is very helpful to researchers working with several different data cultures. Educational and public users may not be familiar with spatial gridding conventions. It is easy to picture the confusion that different map projections create. Students used to seeing how large Greenland is with respect to Africa on a Mercator projection may be shocked at how much Greenland shrinks when it is displayed on an equal-area projection. Both of these communities need simple narratives that can help users locate map features.

34. We can find examples of this diversity even within a single disciplinary community. The International Satellite Cloud Climatology Project (ISCCP) uses an 'igloo-like' gridding scheme that is approximately equal-area. ISCCP starts numbering its boxes at the South Pole. ERBE uses an equal-angle gridding whose box numbering starts at the North Pole. When the author wants to compare ERBE fluxes with ISCCP clouds, he has to be careful about transforming from one grid to another. Since he has lived with these two data sets for a long time, the author simply expects to take some time in dealing with the underlying grids and indexing approaches. Other users are not so patient.

35. Getting the communities to establish a common and well-documented agreement on these underlying conventions is certain to exceed the patience of Job (and the travel budget of even the Defense Department). The author recalls discussions on establishing a common grid for EOS. These talks extended over a year and a half without really reaching a consensus that was enthusiastically supported by the EOS scientific communities. It is probably more realistic to work toward getting each community to document its conventions. Then, the archival community can at least point users to this documentation and, perhaps, provide software to translate from one convention to another.

36. Reaching a consensus on the underlying spatial and temporal structures and on the software services that use them is more difficult for scientific data than it seems on the surface. The sampling processes that create these data vary from instrument to instrument and mission to mission. Sophisticated users may want to enhance the resolution of digital data or remap it to conform to conventions not in the original data. Interdisciplinary users

combining data from different sensor systems and different missions also need to resample data so they can put data in their own spatial and temporal structure. These users face particularly difficult problems in obtaining consistent, reliable, and quantitative documentation from multiple sources about the PSF, calibration, and algorithmic basis for the data products they use in their research.

37. It isn't easy to supply software to scientific data users. The number of scientific software users is much smaller than the number of users for utilities like word processors. Thus, vendors have a smaller user base over which they can amortize their development effort. In addition, the scientific community is 'tribal' in nature, with each tribe having its own 'data religion'. Software vendors are left to face diverse, warring communities that create niche markets. Standards may help, but scientists don't like to work on standards. Developing standards takes time away from scientific research. 'Shareware' developed by researchers may help other users, but scientists are often hard-pressed to support a variety of hardware and software products. If data centers agree to distribute such software, they face difficult resource allocation decisions as well as problems with user expectations and liability.

3.3 User Spatial 'Objects'

38. There are other differences between the data worlds of producers and of users. Producers think in terms of large blocks of data. The author believes that many data users think in terms of objects that come from different spatial classes:

- Fixed points in space (cities, islands)
- Fixed regions in space (continents, ecosystems, snowfields, and short-term regions associated with field experiments)
- Moving, irregularly shaped objects (storms, dust clouds, smoke plumes, fires)

Figure 5 illustrates these classes. It also places the classes in the context of the kinds of data structures producers are likely to use.

39. Cities and islands are relatively easy to identify in longitude and latitude. So are fixed regions, such as continents or field experiments. Identifying ecosystems and snowfields requires expert help. For a definitive decision on whether a remote sensing feature is a snowfield or a glacier, consult a glaciologist. Scientists certainly need to help identify moving targets such as storms, dust clouds, smoke plumes, and fires. Distinguishing between clouds and smoke is not a trivial problem.

40. There is considerable interest in 'data mining' techniques that identify 'interesting objects' in data. Marketing or financial engineering may benefit from such techniques. However, their applicability to scientific data is unproven. Scientific researchers place great importance on the heritage and physical basis of the algorithms they use to derive results. This community will expect object identification algorithms to be understandable and repeatable. Feature identification algorithms should separate one class of objects from another with well-understood uncertainties. Proposals to apply non-peer-reviewed 'data mining' techniques to remotely sensed data will probably not achieve scientific credibility.

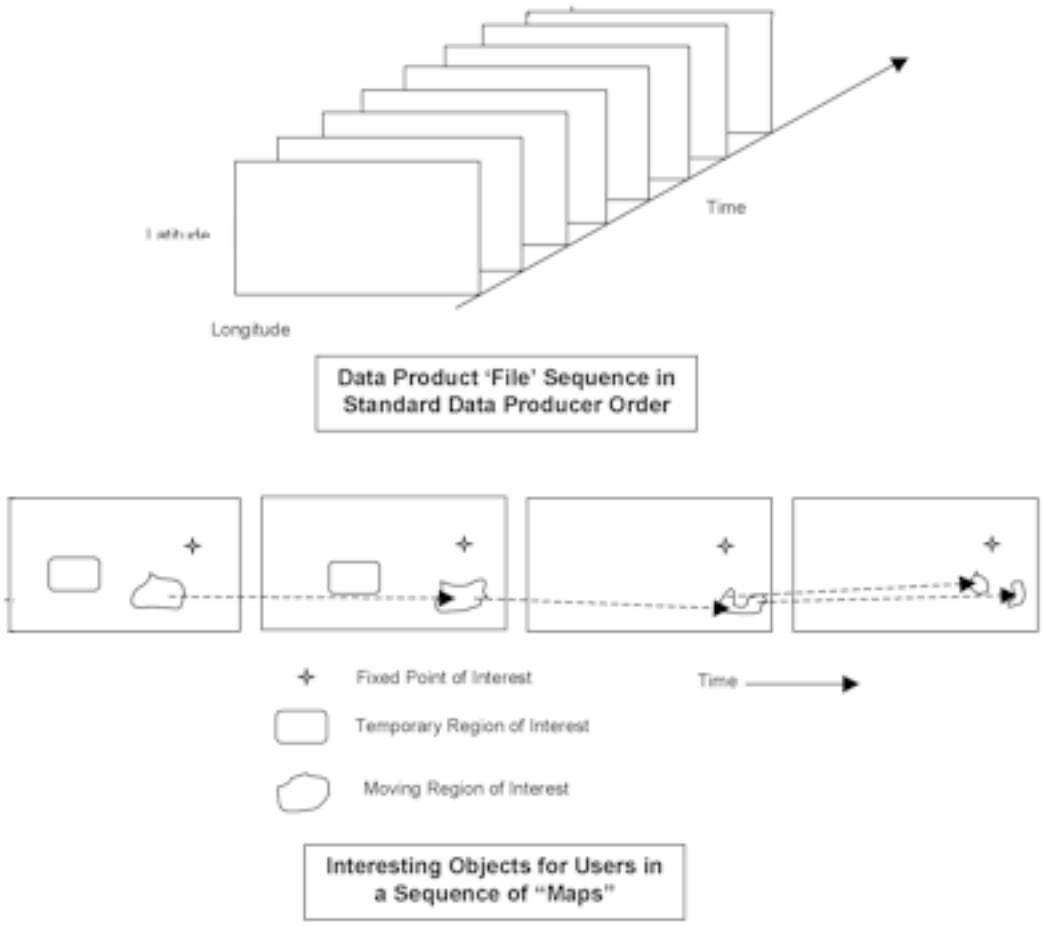


Figure 5. Spatial and Temporal Structures for User Objects. The upper sequence of rectangles illustrates a time sequence of 'global maps', each of which is contained in a single file. This is the 'standard order' in which a data producer has inserted the files into the data set. The lower sequence of rectangles illustrates a time sequence of 'global maps' in which the first user is interested in a fixed target, the second is interested in a region for a brief period of time, and the third is interested in a moving target that changes shape over time. The dashed arrows for the moving target are intended to help the reader see that the object is the same from one map to the next.

3.4 Data Search Services

41. Each of the user spatial classes we described creates challenges for a data center trying to help users obtain data. Data centers need several approaches to help users find data for each of these object classes. Two approaches come to mind immediately:

- One approach is to make the spatial boundaries of the data in each file visible in the file metadata. The user searches this metadata to find files whose data boundaries include the target of interest. The data center software then delivers the files to him.
- A second approach is to create secondary indexes for commonly needed objects. The user interacts with software that understands the indexes. From this interaction, he selects the objects in which he is interested. Then, other software retrieves the data for those objects and distributes it to him.

The first approach delivers files; the second delivers data for objects. Between these two approaches there is a

considerable spectrum of search services and delivery mechanisms.

42. Let us broaden the discussion to the general problem of helping users find data, that is of providing search services. There appear to be five primary approaches to providing these services:

- From an inventory that contains lists of the data products, data sets, data set versions, and files within data set versions
- From keyword lists that link parameters to data products
- From statistical summaries of the data in individual 'files':
- Aggregation statistics
- Structure preserving summaries
- From documentation
- With secondary indexes

Let us briefly consider each of these methods.

3.4.1 Inventory Search

43. In the first approach to finding data, the data center maintains a user-accessible representation of the tree structure we displayed in figure 4. Users traverse that tree structure to find the files they want. Data centers can provide this search tree in several ways. Small collections might have HTML presentations that would allow users to traverse from page to page until they located the files that interested them. Larger collections of data files will use databases to store this kind of information. Because the tree structure can be portrayed as an outline, this approach to finding data is very similar to using a 'Table of Contents' for a book. In this metaphor, the files correspond to pages of text. Data Products correspond to chapters, Data Sets to sections, and Data Set Versions to subsections of text.

3.4.2 Parameter (Keyword) Search

44. In the second approach, users locate appropriate data products by searching through tables that list parameters contained in the data product files. These tables can use the same implementation mechanisms we just discussed for using the tree structure of file collections. A parameter list that points to files or file collections is similar to an index for a printed text. However, the metaphor isn't exact. The parameters in data products are nearly identical from one data set to another. If we take this metaphor literally, it would apply to books that used different words in different chapters, but in which the sections, subsections, and pages within a chapter had the same words. As we suggest below, developing a list of parameters in files is probably easiest to do from documentation.

45. Parameter lists also need to incorporate *aliases* and related kinds of 'pointers'. A user who wants to find data that contain the Earth's 'radiation *balance*' may not think to look for data containing the Earth's 'radiation *budget*', even though these two terms are essentially identical. To further complicate the job that data centers undertake, terminology evolves as the communities of discourse evolve. An index for the *Philosophical Transactions of the Royal Society* in 1700 would contain a vastly different set of words and phrases than one for that publication in 2000. Data centers need to find ways of identifying terminology that their data producers use

when they are generating their products. Then, the data centers need to periodically review the evolution of that terminology.

3.4.3 Metadata Searches

46. In the third search method, users examine statistical summaries of the data in individual files. Data centers seem to associate the term *metadata* with these summary values. We can distinguish between two different kinds of statistical summary data:

- *Aggregation statistics* that provide summary values for the various fields in a file, and
- *Structure preserving summaries*, such as browse images, that sample the data in a file and preserve the underlying spatial or temporal order of the data

Both of these kinds of metadata present issues for data centers.

3.4.3.1 Aggregated Statistical Values

47. Aggregate summaries of the data in a file may be useful in searching for particular files if the summary values differ significantly from one file to another. Where data files contain distinctly different samplings, widely separated in space, such summaries may be quite useful. For example, the EOS ASTER data are images about 60 km on a side. A user might well be able to use the average of the ratio of two spectral bands over the whole image to separate files from one another. CERES ES8 files provide a counter-example. These files contain data from a single satellite for an entire day. The average fraction of the data identified as ‘Overcast’ within each ES8 file is so stable that searching on this fraction will provide no useful distinction between one file and another. Data users still need some sense of the semantics of these statistics in order to use them wisely.

48. The author includes spatial and temporal boundaries in the aggregate summary metadata. When data searches use these boundaries, the boundary representation can be important in the data center’s response to user queries. For example, in dealing with satellite data, it may be more efficient to use a coordinate system oriented with respect to the orbit path and distance along that path. Such an approach is similar to the ‘Row-Path’ representation Landsat uses. On the one hand, this coordinate system appears to make it easy to find coincident data from several instruments on the same satellite. On the other hand, an orbital coordinate system is more difficult for users to interact with.

49. To simplify the interaction between users and a data center, a system designer might be tempted to use an Earth-fixed bounding rectangle to summarize the spatial limits of the data in a file. Certainly an Earth-fixed coordinate system using longitude and latitude is relatively easy to relate to other reference material. However, that referential simplicity carries a price. Figure 6 illustrates some of the geometric problems that an Earth-fixed, bounding rectangle approach to spatial search may have. Finding an appropriate match between the data structure and the search mechanism is one of the ‘semantic’ problems that face data producers, data users, and data centers.

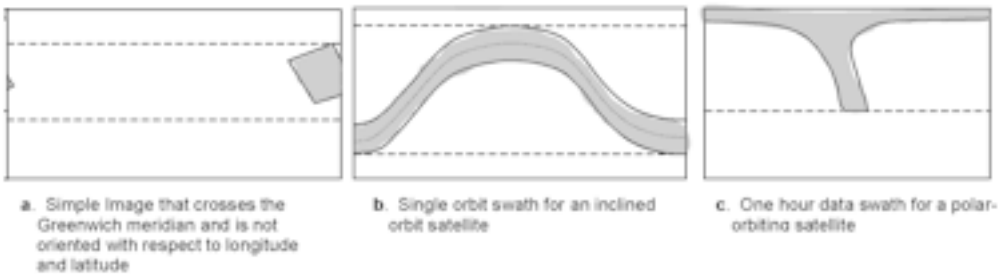


Figure 6. Orbital Geometry Complications in an Earth-Fixed (Longitude–Latitude) Representation. The shaded geometry in each of these maps represents the portion of the Earth sampled in the data file. The dotted lines represent the bounding rectangles that we might use in an Earth-fixed coordinate system to help users conduct a spatial search. For each of these geometries, the Earth-fixed bounding rectangle would produce a 'false positive' query response for many user locations. The ratio of shaded area to total area in the bonding box gives the fraction of 'false positives'.

3.4.3.2 Structure Preserving Summaries

50. Structure preserving summaries raise many of the same issues that statistical aggregate metadata does for producers, users, and data centers. Browse images are a classic example of this kind of metadata. In generating a browse image, a data producer will sample the original image with a lower frequency than the original image used. As an example, a data producer might choose to place every tenth pixel of every tenth line in the browse image. If the original image were 2000 pixels by 2000 pixels, the browse image would be 200 pixels by 200 pixels. The data volume of the browse image is thereby reduced by a factor of 10,000 from the original image.

51. Browse images (and similar summary search structures) are useful, but they do not solve all search issues. Users still have to be very careful about the characteristics of these summaries. We need to understand the possible difficulties users face when they try to relate summaries of files that have vastly different spatial and temporal resolutions.

52. A MODIS image and a CERES ES8 file provide a useful example of this difficulty. The MODIS image includes about two minutes of data and covers an area about 2000 km by 2000 km with a resolution of 1 km. Let us assume that the MODIS team generates a browse image by taking every tenth pixel of every tenth scan line. In other words, the browse image contains data with a resolution of 20 km by 20 km. The CERES ES8 file includes twenty-four hours of data and covers the whole globe with an average resolution of about 30 to 50 km. One immediate problem with a summary of the ES8 file that preserves static spatial structure is that the data covers most areas of the Earth twice in one file. Suppose we build a grid of 100-km regions to summarize the ES8 data product. An average region will have two ES8 data values – one during the day and one during the night. For the sake of argument, we create a browse image for this spatial grid by accepting the last non-zero value that was observed.

53. Now suppose a user wanted to compare a file of ES8 data with a MODIS image. Would the browse images of each file be useful? The MODIS browse image has a spatial resolution comparable with the original resolution of the CERES ES8 data. The ES8 browse image would have about four hundred regions within the MODIS browse image's bounding rectangle. However, some of the regions would have data from the first satellite overpass, others would have data from the second. Comparing these browse images might – or might not be useful.

3.4.4 Documentation Search

54. In the fourth approach, users search for data based on documentation provided by the data producers and the data centers. In the Earth science communities served by EOSDIS, two forms of documentation have become standard: the Algorithm Theoretical Basis Document (ATBD) and the ‘Guide’ documents that the EOS Distributed Active Archive Centers (DAACs) provide. The ATBD’s provide moderately detailed descriptions of the algorithms, the input data, the output data, and important intermediate data products for each EOS investigation. The EOS data producers write the ATBD’s for their own investigation. After they are written, the EOS Project arranges for peer review and discussion of these documents. In the end, the ATBD’s appear as WWW documents that can facilitate data use by individuals who are quite unfamiliar with the background of a given investigation. The EOS Guide documents are brief summaries of data products. The DAACs typically write the Guides for WWW access by the user community. Both of these documents are examples of background material that data producers and data centers provide to help users locate and productively interact with data.

3.4.5 Secondary Index Search

55. With the fifth method, users traverse secondary indexing structures to find data objects that interest them. As we commented previously, data producers tend to concentrate on rather uniform ‘chunks’ of data to ease their production software. When a data producer identifies ‘objects’ in his data, they are likely to be closely tied to the original intent of the investigation. For example, ERBE and CERES identify atmospheric columns that are clear and distinguish these columns from those that have some cloud. It is relatively straightforward to extract clear-sky data from either project’s data products because the notion of ‘clear-sky’ is already embedded in the data. However, the ERBE and CERES data producers do not identify "storm systems" or "hurricanes" in their products. Later investigators have to supply algorithms to locate these new objects and identify the data values in the original data product files that belong to them.

56. Building secondary indexes appears to be a very cost-effective way of increasing the value of data in data centers. It does not involve massive investments in new instruments and launch vehicles. On the other hand, this approach does require access to data sets that may be quite voluminous. It also requires developing the software to identify and retrieve the objects.

3.5 Print Technology and Hypertext

57. The five search methods we have just discussed use "metadata" to help users search. As figure 7 suggests, much of our current thinking about metadata appears to have structural similarities to the search mechanisms we are familiar with from print technology. As Landow [1997] points out, print technology has taken about four hundred years to evolve reasonably standard forms of annotation that guide readers to information they need. This navigation technology includes almost unnoticeable devices such as spaces between words and periods at the end of sentences. It includes indentation or white space between paragraphs. This technology also includes page numbers, section and subsection headings, or even "chapter and verse" references. At the level identified in figure 7, these devices include Tables of Contents and Indexes.

58. Figure 8 suggests two extensions to these aids that move us in the direction of employing ‘hypertext’

technology. Such technology would allow the entire data using community to create entirely new ways of exploring and interacting with scientific data. It stands in marked contrast with ‘print’ technology. The latter creates an expectation that there is a unique sequence that properly orders data. Users should strictly follow this sequence when they want to interact with data. A ‘hypertext’ view suggests that there are many possible sequences in which users can traverse a site to find meaning.

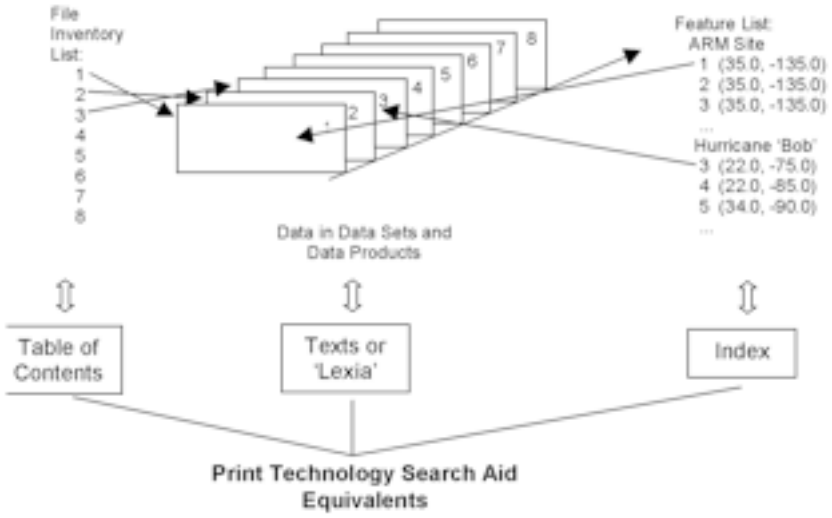


Figure 7. ‘Print’ Technology Search Data Structures. In a non-fiction-printed document, we expect two kinds of data structures to help us navigate through the text. This figure identifies some analogues for the data structures we have identified: a searchable ‘file inventory’ serves as the equivalent to a ‘table of contents’, a ‘list of features’ serves as the equivalent to an ‘index’. In practice, the inventory structure will have more hierarchical layers corresponding to the hierarchy we introduced with ‘files’, ‘data set versions’, ‘data sets’, and ‘data products’. Likewise, there may be several lists that provide random access suggestions to the data itself. These random access items generally are grouped into ‘metadata’.

59. Figure 8 illustrates two ‘hypertext’ traversal mechanisms. At the lower right, we illustrate the secondary index search mechanism. Near the top, we suggest that data producers might embed pointers from one file to another. In a truly object-oriented approach to data, the pointers could also contain references to functions users could activate to interact further with it.

60. As we explore below, this view of data and its access mechanisms opens up much more fluid possibilities than we might expect from previous visions of data centers and data archives. The new vision also creates new problems. Solving them and reducing the experimentation we experience in the WWW to a body of accepted, standard, and useful practice constitutes part of the interesting journey we have embarked upon as a scientific community.

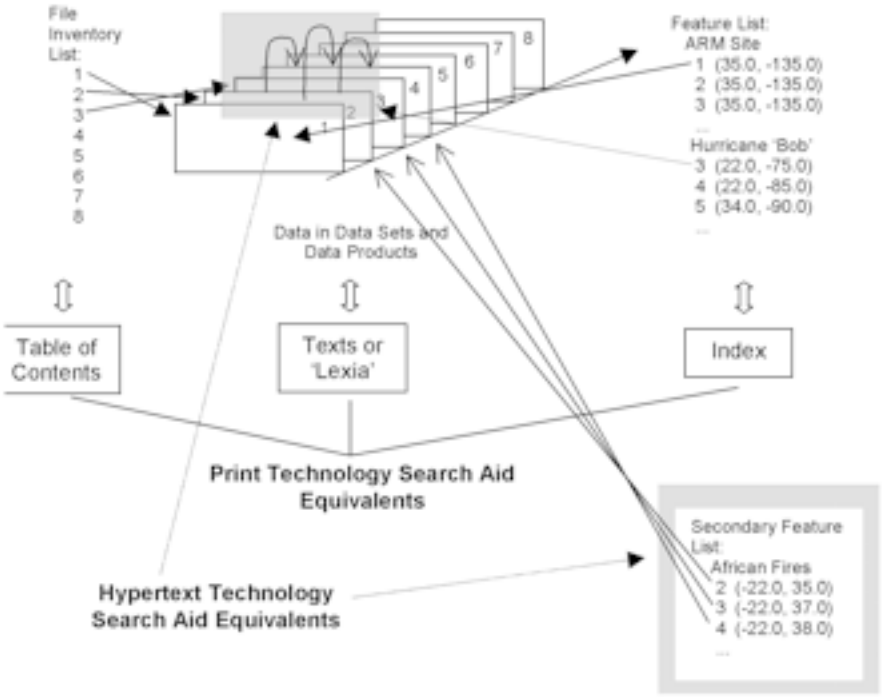


Figure 8. 'Hypertext' Technology Search Data Structures. This figure illustrates 'hypertext' styles of linkages directly from one data feature to another. In a fully object-oriented approach to data, these pointers could become active elements that invoke 'methods' to create particular kinds of responses from systems that would recognize the appropriate semantics of the methods. The figure is more conservative in illustrating only passive links between features in the data, or in adding secondary indices to the structures created by the data producer.

61. When an investigator develops a method of identifying interesting objects in one of these data sets, it is useful to think about how that method might be extended to identify similar objects in the other data sets. For example, suppose an investigator develops a method of identifying "storms" in the ERBE data and can extend that method to the CERES data. To take maximum advantage of LIS information, our investigator needs to be able to identify "storms" in the LIS data set and to identify how these two views of the underlying phenomena complement each other.

62. There is no free lunch in the digital archive world. The next section explores some of the burdens the hypertext approach to data brings with it. Thinking that all we have to do is to build secondary indexes is much too simple.

63. Figure 8 is relatively 'tame'. More interesting possibilities open when we consider extending the objects users can interact with from one data set to many. Figure 9 illustrates the underlying topology that this possible extension suggests. In the foreground, we have a data set obtained early in our remote sensing history. Later, the community makes a new data set with instruments that share a common sampling and measurement capability. As researchers work with the later data set, they realize that another data set in the same period provide significant additional information. For example, we might consider the TOA fluxes from ERBE as an early data set. CERES provides a continuation of that data into the EOS era. In the new era, Lightning Imaging Sensor data on TRMM provide new insights into where clouds have ice. The LIS data may add considerably to the value of the CERES data.

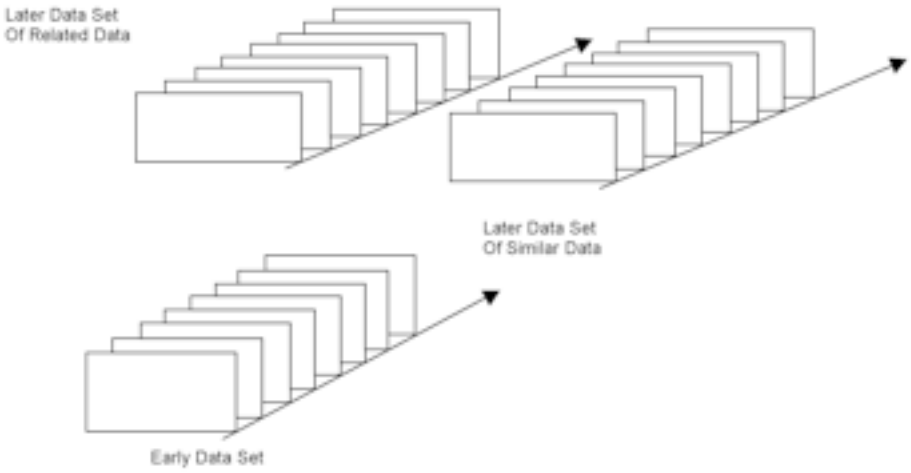


Figure 9. Configuration Management Relationships among Data Collections. This figure illustrates three data sets. The one in the foreground was collected early in the history of this data producer community. The second, in the background right is similar to the first in general characteristics, but has different spatial and spectral sampling. The third, in the left background is a related data set obtained with different sensors and containing different physical quantities.

3.6 Inter-Data Collection Configuration Management Issues

64. Configuration management issues connected with scientific data sets over long time horizons are easy to identify in figure 9:

- Sensor calibration issues – How do we document and accommodate instrument calibrations that vary across gaps of years and with changes in detector technology?
- Instrument change issues – How do we deal with major and minor changes, even with instruments that are intended to collect the same scientific data? For example, how should we deal with changes in sensor Point Spread Functions, spectral sensitivity, and scan patterns?
- Sampling issues – How do we deal with major and minor changes in sampling? For example, under what conditions are data from a Sun-synchronous orbit equivalent to data from a precessing orbit? What consistency checks are possible when one country's geostationary imager goes from top to bottom and another goes from bottom to top? How should we treat scan start times for images, when one country starts a half-hour scan fifteen minutes ahead of a similar scan pattern on another country's imager? Should we worry about altitude changes from one satellite instrument to another when the angular field-of-view is held constant?

65. In each of these system configuration management issues, we can see challenges for digital archives – not just in dealing with changes in the format of the data stream, but with the scientific (or semantic) content of the data. Although we present these issues in the section on the user perspective on scientific data, they look forward to the next section, in which we take the data center's perspective.

4 An Archive View

66. The previous two sections provided a perspective on data production and search from the standpoint of communities whose members lie outside of data and archival centers. In this section, we want to sharpen our perception of some of the issues these organizations face from the inside. We put these perceptions into the context of an ‘Open Archival Information System (OAIS)’, where there is a strong emphasis on interactions between organizations involved in these kinds of efforts.

4.1 Producer Data Ingest and Production

67. We begin our discussion of the data center view by thinking about the interface with the data producers. An archive center certainly has a strong interest in reducing the differences in the unique infrastructure they have to build for each producer. There are enough difficulties in dealing with the differences in scientific data content to keep data centers busy for a long time. It would seem useful to develop standard interface templates for the kind of material data centers need if they are going to produce data or if they are going to accept it from producers. Certainly data centers want to allow users to search for the data and to order it. The CCSDS *Reference Model for an Open Archival Information System (OAIS)* [1998] provides the beginning of such a standards-based approach. This Reference Model concentrates on developing a framework that will allow data centers to identify common ‘packagings’ that Earth science data producers can use to develop data collections that archives can ingest.

68. The Reference Model is still silent regarding some of the issues that we have raised about what data producers need to understand. Perhaps the most important of these is the issue of preserving the semantic (i.e., the scientific) content of Earth science data. While data formats are important, they are considerably more transient that the underlying sampling structure that is unique to scientific data. It would appear that the data centers have a responsibility for bringing these semi-conscious structures into clear visibility. At the same time, data producers cannot be bullied into providing what the community needs over the long haul. It is probably not even possible to attain a uniform and consistent set of definitions without some incentive that is more concrete than "the community’s interest and the common good". The most practical approach may be to develop a community consensus on a (hopefully) small set of standards in each discipline that provides data. Over time, the visibility of the standards may evolve towards a community consensus across the whole realm of the Earth sciences. In natural systems, energy is required to create order out of disorder. Data systems exhibit the same requirement.

4.2 User Data Searching and Distribution

69. The data center interface with users is similar to the interface with data producers in its diversity – except that there are more types of users and more of them. Unanimity is not to be expected within the foreseeable future. Indeed, homogeneity in interfaces is probably undesirable because it removes the interfaces from experimentation and regeneration. In other words, we should expect data centers to innovate and even to compete. Diversity is interesting and beneficial in the long run.

70. Technology will continue to ensure some of this useful diversity. Careful examination of the existing data distribution statistics shows a clear pattern:

Many users order small amounts of data; a few users order large data amounts.

We do not expect technology to change these habits in the near future. Media shipments will be with us for a long time.

71. At the same time, we do not expect the user community to stand still. The prevailing paradigm is moving in the direction of using objects. We will comment later on the overall effect of this change in user expectations. Using objects in a hypertext context should radically change the way data centers work. One of the most profound changes appears in the services we discuss in the next subsection: subsetting and supersetting.

4.3 Subsetting and Supersetting

72. In the classic approach EOSDIS has taken to providing data access, users are forced into thinking about ‘tidy’ but large packages of data – the files created by the data producers. If the users are similar to the data producers and are prepared to handle these packages, there is no real problem. On the other hand, if users are thinking within a framework that is different than the one the data producers had, the user’s life may be more complicated.

73. One of the problems users encounter is that the uniform ‘chunking’ of data that is natural to a data producer also creates files that include more data than many users want. If a user wants ERBE longwave fluxes over the Sahara desert, he has to get ERBE longwave fluxes for the whole globe. If a user wants one day of ISCCP cloud data, he has to order a whole month. If a user just wants data from Lake Geneva in Walworth County, WI, he has to order the entire Landsat image that includes Walworth, Green, and Rock counties. It is as if the data providers had never encountered retail butchers, but insisted that data users accept the whole cow and butcher it themselves.

74. To be fair, users often want rather strangely shaped objects. A NASA Headquarters manager once offered to allow the author to take data from the state of Virginia as a subset – and didn’t realize for several years how oddly shaped such data chunks might become. Algorithms to provide a reasonable user selection of subsetted data may not be trivial, particularly since such algorithms need to be highly efficient to be cost effective.

75. As we have already seen, the user object orientation often crosses the boundaries of the data chunks created by data producers. When this happens, users need to be able to create a new collection of data from pieces of existing chunks. In other words, they need to be able to superset subsets.

76. As figure 10 shows, we can view subsetting and supersetting as methods of creating new ‘file’ objects from old ‘file’ objects. In slightly different (more UNIX-like) language, subsetting and supersetting become a filtering operation in a ‘production pipe’. For this to take place, users need to be able to identify the files they need, the subsets they want from each file, and the order in which the final objects appear in the superset file.

77. The performance issue is similar to the performance of ad hoc queries in databases. We need methods of optimizing the query structure to minimize the load on the data center resources. Of course, there is a spectrum of solutions to this problem. At the extreme end lie queries that we could state as the classic "Polar Bear" problem: "Find all the polar bears under the Arctic Ice during a winter when there is an ozone hole and where there is an algal bloom." [The perverse form of this query arises when "Polar Bear", "Arctic", "ice", "winter", "ozone hole", and "algal bloom" are objects that the data system must construct from understanding natural language constructs and the contents of the existing system. A related query can be stated "search through all the data in all the archives to find me something interesting to think about." It appears that these queries are NP-hard. To the author, there are enough scientifically interesting problems that this kind of computer science research is

uninteresting.] Trying to set up systems to solve these queries automatically does not seem useful at this time. The more interesting issues arise when we try to couple human knowledge with computers – to form a synergistic approach to the query optimization problem.

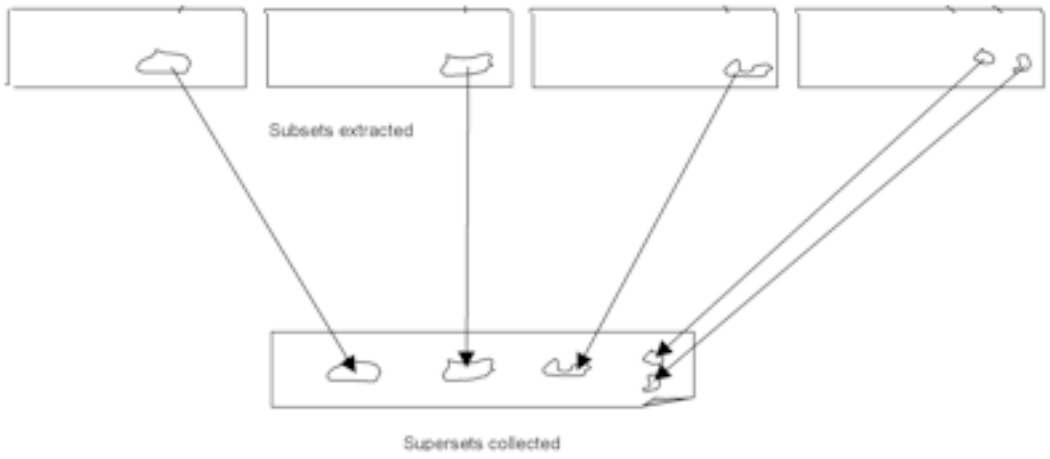


Figure 10. Subsetting and Supersetting as Ways of Overcoming Differences between Producer 'Data Chunking' and User 'Object-Orientation'. The rectangles represent files, such as those we illustrated in figure 5. The subsetting service at a data center extracts 'objects' from these files and places them in a new file containing the superset. We can think of this combined operation as creating a new file by filtering several old ones. If the data producer has carefully incorporated the Annotation Records we identified in figure 2, then the file format of the subsetted files will be identical with the file format of the superset file.

4.4 Semantic Requirements for Data Interchange

78. Both data producers and data users are moving toward an "object-oriented" approach to scientific data. In the long-run, both communities would benefit from a distributed "information architecture". With such an architecture, interdisciplinary investigators could set up automated procedures in their own facilities that would find appropriate data at several different data centers, subset and superset the data files, and extract the results for their own use. Educational users could find lesson plans that automatically extracted sample data and provided software to experiment with. The public could sit down to an automated 'documentary' that extracted and displayed complex images to the accompaniment of audio clips. For now, such a vision appears to be an interesting goal, but one that will require an extraordinary amount of work to achieve. The problem is not merely technological; it is sociological. To be practical, we need to find ways of fostering cost-effective data centers. Almost certainly this means avoiding attempts to build "one-size fits all" systems.

79. To ensure cost effectiveness, it seems reasonable to suggest that interoperability is only required where there is significant interchange of data. For this reason, interoperability between NASA's space science enterprise and the agency's Earth science enterprise probably do not need much more than the ability to exchange lists of parameters and the data products that contain them. Only as communities move toward continual exchange of data do we need to arrange for long-term commitments between institutions.

80. When we look at these institutional interfaces in this light, it is clear that the interfaces are one of the costs of community interoperability. We cannot avoid the fact that useful data interchange between archives requires

common semantic structures and content. Astronomers locate their data by Right Ascension and declination; Earth scientists locate their data by latitude and longitude. We could waste a lot of time and money trying to merge incompatible structures.

5 Tentative Conclusions

81. Where do these views lead us? One perspective suggests that over the long term, scientific data archives will move to an object orientation. In the previous text, we suggested what such object orientation means for Earth science data. Here, we summarize that view. We identify some interesting issues for digital archives. Some of these issues are particular to scientific data; others are common to all kinds of digital data. After raising these issues, we suggest some opportunities to move forward.

5.1 An Object Oriented View of Archive Information Evolution

82. Looking at the architecture of data centers over the next twenty years, several computer science groups envision the possibility of *Cooperative Information Centers*, which are made of heterogeneous data centers that exchange data, e.g., Papazoglou, M. P. and G. Schlageter [1998]. Authors espousing this idea have had enough experience to feel that it will take a long time to implement. On the other hand, this vision does suggest a future that contains individual data centers that cooperate to provide services that are more helpful than any could provide alone. This vision does not require a single, homogeneous approach. It fits well with the structure we have suggested in the previous sections of this paper.

82. First, we do not see a future in which data producers and data users have a single view of what they want from data. We expect that the data producer's 'chunks' and the user's 'objects' will always be part of the data landscape. This dichotomy is probably a permanent feature of scientific data archival.

83. Second, it is also clear that we need a more visible information structure than we have had so far. Long-term archival requirements for scientific data include need for documenting

- Underlying spatial and temporal sampling structures
- Data structures of archive holdings
- Algorithmic basis for producing data
- Production topology

84. The underlying, semantic structure has not been as visible as the formatting issue for scientific data. However, getting this structure documented is probably more important in the long run. Likewise, it is difficult to see how we can make progress in providing effective data services without paying considerably more attention to the data structure of archival holdings. The file and metadata structures determine how long it takes to respond to user requests. We have made some progress in documenting the scientific algorithms for data production. Because these algorithms are very discipline dependent, it is difficult to foresee useful standardization of these key data production tools. In addition, when we put algorithms into a data production environment, we have to become much more systematic about documenting the production topology, the connectivity between 'jobs' and 'files' that determine the genealogy of scientific data in an archive.

85. Third, long-term data access and retrieval requirements for scientific data need mechanisms that allow

- User searches by objects as well as spatial and temporal intervals
- Development of alternative data groupings and retrieval services after primary data production
- Development of methods of bridging heterogeneous data collections (text annotation versus hypertext annotation)

5.2 Scientific Data Archival Issues

86. We can list the issues these considerations raise as follows:

- What are the mechanisms for standardizing sampling structures – in data formats and in documentation, including the semantic content of query structures and representation of bad data?
- What are the mechanisms for describing data production topology, history, and collections? – Can we develop standard collection nomenclature and algorithms for deriving versioning schema?
- What standards do we need for describing data collection structures, such as tree level names and relationship to hypertext documentation?
- What mechanisms do we have for evolving search mechanisms, including
    - Keyword entries and keyword aliases
    - Metadata structures, including both statistical aggregates of data 'files' and structure preserving summaries of these entities
    - Secondary indexing approaches, particularly those provided by third parties
- What approaches should we take to providing homogeneous views of heterogeneous data collections, particularly historical data in which we want to identify equivalent 'features', including such issues as
    - Calibration differences
    - Spectral band sampling and orbital sampling differences
    - Spatial resolution differences
    - Algorithm differences

5.3 A Perspective on the Future of Digital Archives for Scientific Data

87. The evolution of the scientific data environment will change the needs of data producers and data users in a number of interesting ways:

- Data producers may move toward a more 'database-centric' view, but only after performance issues are resolved (move from MFLOPS to TBS)
- Data users may move toward a more 'long-term object' view, but probably from a disciplinary basis
- 'Hypertext' view of connectivity and automated search approaches will influence the long-term evolution of scientific data production and user searches

88. When we think about these changes in the communities that use data archives, we may also expect these institutions to evolve as well. It is not clear at the moment whether it is more effective to hire Ph.D. computer scientists at the data centers or to raise the computer science expertise of the data producers and data users. Regardless of the approach, the environment of continuous change that surrounds information technology will be

with us indefinitely. The entire conception of scientific data archives may change from ‘producer file collections’ and ‘write-once, read-none deep archives’ into ‘communities of data scholars’. It is these communities that will have to carry the responsibility for continually translating the past record of what we have observed into useful information about the future.

6 References

Consultative Committee on Space Data Systems Standards (CCSDS), 1998: *Reference Model for an Open Archival Information System (OAIS)*, **CCSDS 650.0-W-3.0**, April 15, 1998.

Landow, G. P., 1997: *Hypertext 2.0, Being a revised, amplified edition of Hypertext: The Convergence of Contemporary Critical Theory and Technology*, The Johns Hopkins University Press, Baltimore, MD, 353 pp.

Papazoglou, M. P. and G. Schlageter, 1998: *Cooperative Information Systems: Trends and Directions*, Academic Press, San Diego, CA, 369 pp.

Table of Contents for this page

Index for this page